CLAIMS

What is claimed is:

1. An apparatus for searching for a base sequence, which searches for a similar base sequence, which has same length, and is similar to a base sequence to be inputted, using an index, which is for searching a database storing a gene base sequence indicating gene information, and is for searching for a position in said gene base sequence, at which a base sequence having a predetermined length appears, comprising:

an input unit for base sequence, which inputs a base sequence having a length longer than said predetermined length;

an input unit for hamming distance, which inputs hamming distance, which indicates number of bases to be substituted to mismatching bases;

a specifying unit, which specifies two different partial sequences, which are partial sequences of said inputted base sequence and have said predetermined length, and the other portion;

an assignment unit, which distributes and assigns the hamming distance inputted by said input unit for hamming distance to the partial sequence and to the other portion specified by said specifying unit;

a selection unit, which selects the partial sequence having a non-larger total number of substituted base sequences generated by substitution for said partial sequence, in which the bases of the number indicated by the hamming distance assigned by said assignment unit are substituted

to the mismatching bases, from the two partial sequences specified by said specifying unit;

a generation unit for substituted base sequence, which generates a substituted base sequence, which has the hamming distance assigned by said assignment unit, for the partial sequence selected by said selector; and

a search unit, which carries out a search using said index and the substituted base sequence generated by said generator for substituted base sequence as a search key.

2. The apparatus for searching for a base sequence according to Claim 1, wherein said specifying unit comprising:

a first specifying means, in which, if number of bases of the base sequence inputted by said input unit for a base sequence is equal to or less than twice of said predetermined length, one end of one partial sequence of said two partial sequences is conformed to one end of said inputted base sequence, and one end of another partial sequence of said two partial sequences is conformed to another end of said inputted base sequence, so that the other portion does not exist and is not specified.

3. The apparatus for searching for a base sequence according to Claim 1 or 2, wherein said specifying unit comprising:

a second specifying means, in which, if number of bases of the base sequence inputted by said input unit for base sequence is more than twice of said predetermined length, said two partial sequences do not overlap with each other, and said two partial sequences

are specified.

4. The apparatus for searching for a base sequence according to any one of Claims 1 to 3, comprising:

an acquisition unit for similar base sequence candidate, which acquires a similar base sequence candidate, which is a base sequence including said substituted base sequence and appearing in a gene base sequence, based on the search result by said search unit; and

a determination unit, which determines whether the hamming distance between the similar base sequence candidate acquired by said acquisition unit for similar base sequence candidate and said inputted base sequence is equal to or less than the hamming distance inputted by said input unit for hamming distance.

5. The apparatus for searching for a base sequence according to Claim 4, comprising:

an input unit for mismatching base pair, which specifies mismatching base pair, wherein the search unit carries out searching and computing the hamming distance based on the base pair inputted by the input unit for mismatching base pair.

6. The apparatus for searching for a base sequence according to Claim 4 or 5, comprising:

an input unit for distribution of matches, which inputs distribution information indicating distribution of matches between the corresponding bases in the base sequence inputted by said input unit for base sequence and the similar base sequence, wherein

said determination unit comprises,

a determination means for distribution, which determines whether the distribution information inputted by said input unit for distribution of matches has been fulfilled.

7. The apparatus for searching for a base sequence according to Claim 6, wherein

the distribution information inputted by said input unit for distribution of matches is lower limit of length of successive matching bases between the base sequence and the similar base sequence.

8. The apparatus for searching for a base sequence according to any one of Claims 1 to 7, wherein

the length of the base sequence inputted by said input unit for base sequence is 15 to 60, and said predetermined length is 11 to 14.

9. A method for searching for a base sequence, which searches for a similar base sequence, which has same length, and is similar to base sequence to be inputted, by using an index, which is for searching a database storing a gene base sequence indicating gene information, and is for searching for a position in said gene base sequence, at which a base sequence having a predetermined length appears, comprising:

a step of inputting base sequence, which inputs a base sequence having a length longer than said predetermined length;

a step of inputting hamming distance, which inputs hamming distance, which indicates number of mismatching bases to be substituted;

a step of specifying, which specifies two different partial sequences, which are partial sequences of said inputted base sequence, and have said predetermined length, and the other portion;

a step of assigning, which distributes and assigns the hamming distance inputted by said input unit for hamming distance to the partial sequence and to the other portion specified by said specifying unit;

a step of selecting, which selects the partial sequence having a non-larger total number of substituted base sequences generated by substitution for said partial sequence, in which the bases of the number indicated by the hamming distance assigned by said assignment unit are substituted to the mismatching bases, from the two partial sequences specified by said specifying unit;

a step of generating substituted base sequence, which generates a substituted base sequence, which has the hamming distance assigned by said assignment unit, for the partial sequence selected by said selector; and

a step of searching, which carries out a search using said index and the substituted base sequence generated by said generator for substituted base sequence as a search key.

10. An apparatus for searching for a character string, which searches for a similar character string, which has same length, and is similar to a character string to be inputted, by using an index, which is for searching a database storing a character string, in which alphabets are arranged

one-dimensionally, and is for searching for a position in said character string stored in said database, at which a character string having a predetermined length appears, comprising:

an input unit for character string, which inputs a character string having a length longer than said predetermined length;

an input unit for hamming distance, which inputs hamming distance, which indicates number of alphabets to be substituted to mismatching alphabets;

a specifying unit, which specifies two different partial character strings, which are partial character strings of said inputted character string and have said predetermined length, and the other portion;

an assignment unit, which distributes and assigns the hamming distance inputted by said input unit for hamming distance to the partial character string and to the other portion specified by said specifying unit;

a selection unit, which selects the partial character string having a non-larger total number of substituted character strings generated by substitution for said partial character string, in which the alphabets of the number indicated by the hamming distance assigned by said assignment unit are substituted to the mismatching alphabets, from the two partial character strings specified by said specifying unit;

a generation unit for substituted character string, which generates a substituted character string, which has the hamming distance assigned by said assignment unit, for the partial character string selected by said selection unit; and

a search unit, which carries out a search using said index and the substituted character

string generated by said generation unit for substituted character string as a search key.

11. The apparatus for searching for character string according to Claim 10, wherein said character string is a peptide sequence.

12. The apparatus for searching for a base sequence according to any one of Claims 1 to 8, comprising:

a storage unit for repeated sequence, which stores base sequence of said predetermined length appearing repeatedly in the gene base sequence;

a storage unit for repeated sequence information, which stores repeated sequence information, in which the base sequence stored by said storage unit for repeated sequence is correlated with a position in said gene base sequence, at which the base sequence appears, wherein said search unit comprises,

a determination means for repeated sequence, which determines whether said substituted base sequence is stored by said storage unit for repeated sequence, and

a search means for repeated sequence, which carries out search based on the repeated sequence information stored in said storage unit for repeated sequence information, if it is determined by said determination means for repeated sequence that said substituted base sequence is stored by said storage unit for repeated sequence.

13. The apparatus for searching for a base sequence according to any one of Claims 4 to 7, comprising:

a storage unit for similar base sequence, which correlates said inputted base sequence, the hamming distance between said inputted base sequence and a similar base sequence candidate, with the similar base sequence candidate, and stores them, if it is determined by said determination unit that the hamming distance between said inputted base sequence and the similar base sequence candidate acquired by said acquisition unit for similar base sequence candidate is less than or equal to the hamming distance inputted by said input unit for hamming distance.

14. The apparatus for searching for a base sequence according to any one of Claims 4 to 7, comprising:

a computation unit for association rate, which computes association rate between said base sequence inputted by said input unit for base sequence and the similar base sequence candidate acquired by said acquisition unit for similar base sequence candidate, if it is determined by said determination unit that the hamming distance between the similar base sequence candidate acquired by said acquisition unit for similar base sequence candidate and said inputted base sequence is less than or equal to the hamming distance inputted by said input unit for hamming distance.

15. An apparatus for generating ineffective base sequence, comprising:

an acquisition unit for base sequence, which acquires a base sequence having a length longer than said predetermined length;

a generation unit for ineffective substituted base sequence candidate, which generates ineffective substituted base sequence candidate, which is a base sequence acquired by substituting a predetermined number of bases among the bases of the base sequence acquired by said acquisition unit for base sequence;

an input unit for ineffective substituted base sequence candidate, which inputs the ineffective substituted base sequence candidate generated by said generation unit for ineffective substituted base sequence candidate to the apparatus for searching for a base sequence according to Claim 14;

a second input unit for hamming distance, which inputs a predetermined hamming distance to the apparatus for searching for base sequence, to which said input unit for ineffective substituted base sequence candidate has inputted the ineffective substituted base sequence candidate; and

a selection unit, which selects the base sequence having a low association rate from the ineffective substituted base sequence candidates generated by said generation unit for ineffective substituted base sequence candidate, in which the base sequence is acquired by said apparatus for searching for base sequence according to the input by said input unit for ineffective substituted base sequence candidate, and to the input by said second input unit for hamming distance.

16. An alignment apparatus for base sequence, comprising:

a second acquisition unit for base sequence, which acquires a base sequence having a length longer than said predetermined length;

a selection unit for partial base sequence, which selects a partial base sequence, which is a portion of the base sequence acquired by said second acquisition unit for base sequence;

an input unit for partial base sequence, which inputs the partial base sequence selected by said selection unit for partial base sequence to the apparatus for searching for a base sequence according to any one of Claims 4 to 8;

a third input unit for hamming distance, which inputs a predetermined hamming distance to the apparatus for searching for a base sequence, to which said input unit for partial base sequence has inputted the partial base sequence; and

an alignment unit, which aligns the base sequence acquired by said second acquisition unit for a base sequence to said gene base sequence based on the search result acquired by said apparatus for searching for a base sequence according to the input by said input unit for partial base sequence, and to the input by said third input unit for hamming distance.